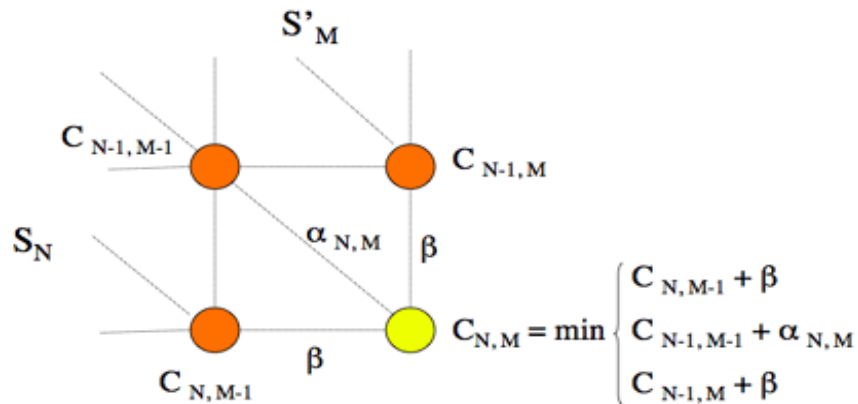


4. Comparaison de séquences

- Comment prédire les fonctions des gènes/protéines ?
- Évolution et similarité de séquences
- Quantifier la similarité de deux séquences
- L'alignement de séquences devient un problème d'optimisation
- Un alignement de séquences vu comme un chemin dans une grille
- Si un chemin est optimal, tous ses chemins partiels sont optimaux
- **Coûts et alignement**
- Un algorithme récursif
- Éviter la récursivité : une version itérative
- Cet algorithme est-il efficace ?

Calcul du coût du dernier nœud

- β : coût d'une insertion « — »
- $\alpha_{N,M}$: coût de substitution de S_N par S'_M
- $\alpha_{N,M}$ est un élément de la matrice `SubstitutionCost [1:4, 1:4]` (pour des séquences d'ADN) ou `[1:20, 1:20]` (pour des séquences protéines)



Séquences et coûts

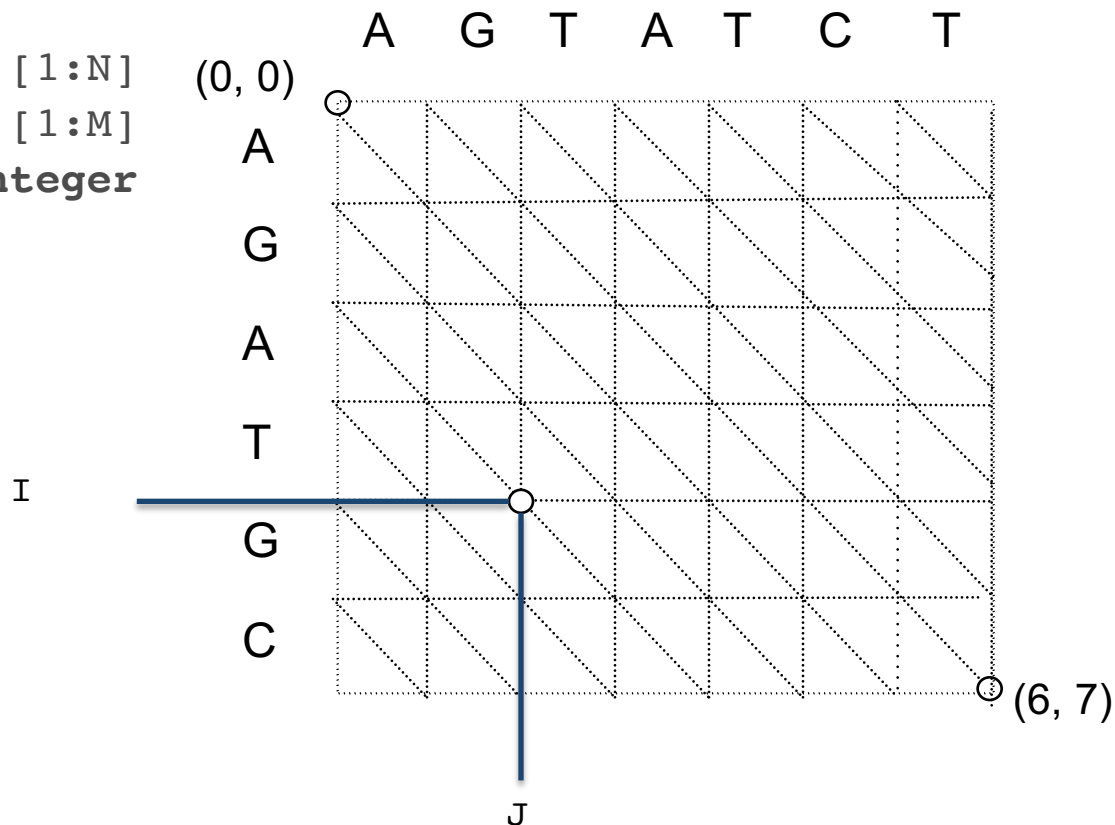
Sequence1: **character string** [1:N]

Sequence2: **character string** [1:M]

Cost: **array** [0:N, 0:M] of **integer**

InsertionCost: **integer**

function SubstitutionCost
(Char1, Char2: **character**)
returns integer



Fonction de calcul du coût de substitution

- Accepte en entrée deux caractères Char1 et Char2
dans l'alphabet de quatre lettres de l'ADN
 $\{A, C, G, T\}$
ou dans l'alphabet de 20 lettres des protéines
 $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, B, Z, X\}$
- Calcule le coût de substitution de Char1 par Char2
- Pour ce faire, recherche dans la **matrice des coûts de substitution**

Matrice des coûts de substitution

- Les valeurs de la matrice reflètent des **considérations biologiques**
 - Par exemple, *considérer différemment les transitions $A \leftrightarrow G$, $C \leftrightarrow T$*
et les transversions $C \leftrightarrow A$, $G \leftrightarrow T$, $A \leftrightarrow T$, $C \leftrightarrow G$
- Plus difficile pour les **matrices 20×20 de substitution des AA**
 - Propriétés physico-chimiques des acides aminés

	A	C	G	T
A	0	1	1	1
C	1	0	1	1
G	1	1	0	1
T	1	1	1	0